SAND2011-2152C

Evaluating Information Visualizations with Working Memory Metrics

Alisa Bandlow, Laura E. Matzen, Kerstan S. Cole, Courtney C. Dornburg, Charles J. Geiseler, John A. Greenfield, Laura A. McNamara and Susan M. Stevens-Adams

Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185 USA
{abandlo, lematze, kscole, ccdornb, cjgiese, jagreen, lamcnam, smsteve}@sandia.gov

Abstract. Information visualization tools are being promoted to aid decision support. These tools assist in the analysis and comprehension of ambiguous and conflicting data sets. Formal evaluations are necessary to demonstrate the effectiveness of visualization tools, yet conducting these studies is difficult. Objective metrics that allow designers to compare the amount of work required for users to operate a particular interface are lacking. This in turn makes it difficult to compare workload across different interfaces, which is problematic for complicated information visualization and visual analytics packages. We believe that measures of working memory load can provide a more objective and consistent way of assessing visualizations and user interfaces across a range of applications. We present initial findings from a study using measures of working memory load to compare the usability of two graph representations.

Keywords: Information visualization, evaluation, cognitive load

1 Introduction

Visual analytics software aims to enhance an individual's ability to make sense of complex data. However, evaluating visual analytics tools is difficult. Good data sets for testing are difficult to obtain. Some lack ground truth while others are sensitive and proprietary. Even when good data sets are available, controlled, experimental testing across tools is difficult when tools support different tasks. Visualization software supports complex tasks that vary across users and domains [7]. Traditional evaluation, including usability studies and controlled experiments, can be "helpful but take significant time and resources"[6]. Moreover, they do not generalize across conditions and contexts, which can lead to costly re-designs for specific data sets and user communities.

2 Cognitive Load Evaluation

We seek evaluation metrics that can be generalized across different types of visual analytics software, data and tasks. Mechanisms of human cognitive processing are consistent across individuals. Since reasoning tasks require substantial cognitive resources, measuring cognitive processing demand can help designers assess the efficacy of visual representations.

Measurements of cognitive load are commonly used in evaluation. However, to our knowledge, the prior uses of cognitive load measures have used subjective questionnaires rather than measures of working memory. For example, the Task Load Index (TLX) questionnaire developed by NASA [2] has been used as one subjective metric for evaluating software tools [5, 8]. However, TLX ratings are subjective and must be combined with application-specific usability metrics, making comparison across software designs tricky and possibly expensive.

A more objective way of assessing cognitive load is to measure working memory, the "theoretical construct that has come to be used in cognitive psychology to refer to the system or mechanism underlying the maintenance of task-relevant information during the performance of a cognitive task" [9].

Working memory approaches can be implemented in a dual-task paradigm requiring completion of two simultaneous tasks. Sternberg tasks are well-validated working memory task frequently used in dual-task studies [10]. A Sternberg task requires participants to remember a distinct set of target items, and then identify them in a string of distractor items. Participants can perform well on the Sternberg task only when their cognitive resources are not consumed by the primary task. Accuracy and reaction times are compared across conditions to assess the relative burden imposed by the primary task. Such secondary task approaches are better at detecting workload than primary task measures alone [1, 4].

Huang et al. [3] suggest that effective visualizations help people concentrate on difficult tasks. Similarly, we suggest that *effective visualizations should minimize the cognitive demands associated with data-driven reasoning*. Difficult-to-use visualizations will consume cognitive resources, minimizing a user's ability to engage in higher-order reasoning. If this is the case, then as individuals are reasoning with a visual representation, performance on a concurrent Sternberg task may indicate if the representation is inducing extraneous cognitive load, beyond that associated with the primary task.

3 Measuring Working Memory Capacity for Graph Evaluation

To assess the feasibility of this approach, we used a within-subjects, dual-task paradigm to assess workload induced by two different graphical representations. Twenty-three participants completed the experiment.

In the primary task, participants reviewed either a traditional vertex-edge graph or a tree-ring graph, and then answered a question about the graph. We presented each participant three versions of each graph, with 20, 40, and 80 elements, for a total of

six graphs. The participants' primary task was to answer six questions about the content of each of the six graphs, for a total of thirty-six questions. Half of the questions were a "good fit" for vertex-edge graph, and half were a "good fit" for the ring graph. The questions that were a "good fit" for the ring graph were a "bad fit" for the vertex-edge graph, and vice versa. The order of the questions and the order in which participants used the graphs were counterbalanced across tasks and across participants. After each question, participants filled out the NASA TLX questionnaire before proceeding. The participants' accuracy, reaction times, and subjective ratings for each question were recorded.

As participants completed the primary task, they completed a concurrent, auditory Sternberg task. We presented a memory set of three random letters before viewing each graph type. A random string of letters was presented over computer speakers at a rate of one letter every two seconds. Participants were instructed to click the mouse button as quickly as possible upon hearing a letter from the memory set. They were told to give more effort to the primary task at the expense of the secondary task.

For the primary task, we hypothesized that participants would take more time to answer "bad fit" questions than "good fit" questions. Secondly, we hypothesized that as the size of the graphs increased, the participants would take more time to respond and would make more errors. Thirdly, we hypothesized that performance differences between the "bad fit" and "good fit" questions would increase as the graph size grew.

We hypothesized that the results of the secondary Sternberg task would mirror the results from the primary task. We predicted that the participants would have longer reaction times and lower accuracy on the Sternberg task when answering the "bad fit" questions and for the larger graphs. We also predicted that the increased graph size would heighten secondary task performance differences between the "bad fit" and "good fit" questions.

4 Graph Evaluation Results and Conclusions

Participants' subjective evaluations indicated the primary task became more difficult for a) larger graphs and b) questions that were a "bad fit" to the graph type. A two-way analysis of variance (ANOVA) for each type of workload assessed by the NASA TLX showed main effects of graph size and question type for the mental demand, temporal demand, effort, and frustration measures (all F=3.90, all p<0.02).

As hypothesized, the participants' performance on the primary task declined as the graphs grew. Decline was greater for questions that were a "bad fit" to the graph type. The average percentages of correct responses are shown in Figure 1 and the average reaction times across conditions are shown in Figure 2. Two-way ANOVAs showed main effects of graph size and question fit and a significant interaction between graph size and question fit for both (p<0.04).

Working Memory (WM) tasks: Our critical prediction was that the working memory measure from the secondary task would mirror the results from the primary task. The participants' responses to the Sternberg task were scored as correct if they

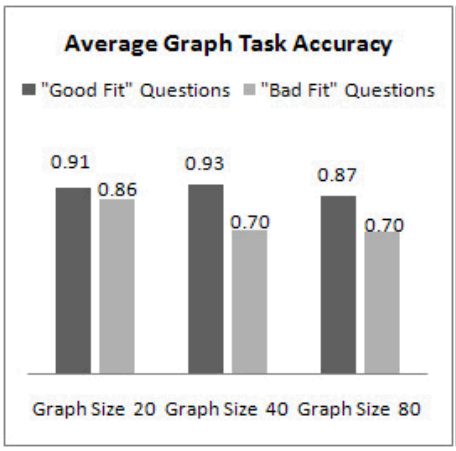

Figure 1

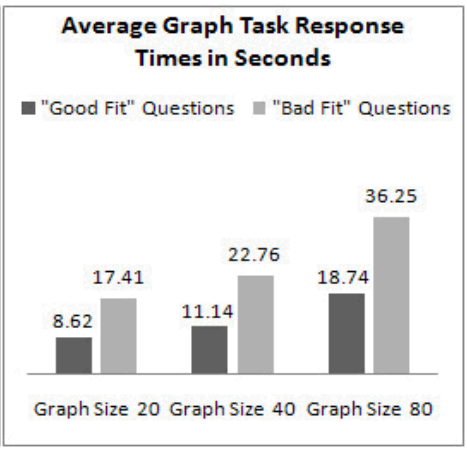

Figure 2

responded to the target letters before the next letter was presented. The participants' hit rates and false alarm rates were used to calculate d' scores, a measure of the participants' ability to discriminate between the target and distractor items. The average d' scores across participants are shown in Figure 3. A two-way ANOVA showed a significant interaction between graph size and question type [F (2, 22) = 4.87, p = 0.01]. This result indicates, as predicted, that the participants' performance on the working memory task reflected the difficulty of the primary task.

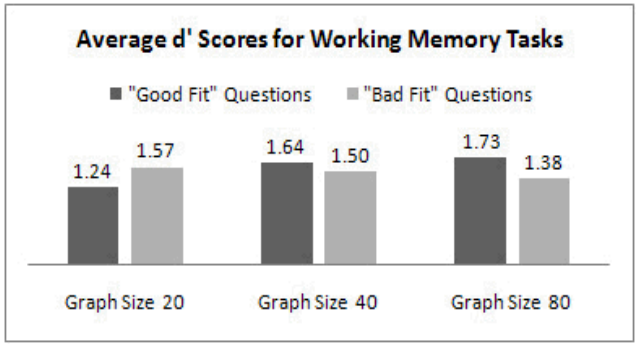

Figure 3

One issue with this evaluation is that the primary task questions were very easy for the size 20 graphs, notably for the "good fit" questions. Participants were often able to answer the questions before any targets were presented in the secondary task, which led to sparse data. In the future, we will ensure that the primary task allows for the presentation of several targets in the working memory task in all conditions.

This study indicates that a secondary working memory task could be useful for evaluating visualizations in cases where it is difficult or impossible to assess primary

task performance. In future work, we plan to extend this method by applying it to more complex visualizations and to user interfaces. We believe that working memory assessments will provide metrics that enable designers to determine if particular design options require more cognitive resources than others for a given task.

Acknowledgements. This work was supported by the Networks Grand Challenge at Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy Company, National Nuclear Security Administration under contract DEAC04-94AL85000.

References

1. Gawron, V.: Human performance, workload, and situational awareness handbook. Taylor & Francis, Boca Raton, FL (2008)
2. Hart, S.G., Staveland, L.E.: Development of a NASA-TLX (Task load index): Results of empirical and theoretical research. In: Hancock, P. and Meshkati, N. (eds.) Human Mental Workload. pp. 139--183. North-Holland, Amsterdam (1988).
3. Huang, W., Eades, P., Hong, S.: Beyond time and error: A cognitive approach to the evaluation of graph drawings. In: Proc. 2008 Conference on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization, pp.1--8. ACM, New York (2008)
4. Meshkati, N., Hancock, P.A., Rahimi, M.: Techniques of mental workload assessment. In: J. Wilson (ed.), Evaluation of Human Work: Practical Ergonomics Methodology. pp. 605—627. Taylor and Francis, London (1989)
5. Morse, E., Steves, M.P., Scholtz, J.: Metrics and methodologies for evaluating technologies for intelligence analysts. In: Proc. Conference on Intelligence Analysis, (2005)
6. Plaisant, C., Fekete, J., Grinstein, G.: Promoting insight-based evaluation of visualizations: From contest to benchmark repository. In: IEEE Transactions on Visualization and Computer Graphics 14. 1, 120--134 (2008)
7. Scholtz, J.: Progess and challenges in evaluating tools for sensemaking. Presented at the ACM CHI conference workshop on sensemaking (2008)
8. Scholtz, J., Morse, E., Steves, M.P.: Evaluation metrics and methodologies for user-centered evaluation of intelligent systems. In: Interacting with Computers 18. pp. 1186-1214 (2006)
9. Shah, P., Miyake, A.: Models of working memory: An introduction. In: A. Miyake and P. Shah (eds.) Models of Working Memory: Mechanisms of Active Maintenance and Executive Control. pp. 1—27. Cambridge University Press, Cambridge, UK (1999)
10. Wierwille, W.W., Eggemeier, F.T.: Recommendations for mental workload measurement in a test and evaluation environment. In: Human Factors. 35, pp. 263--282 (1993)